International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development of a Predictive Risk Model for Recurrence of Chronic Pulmonary Aspergillosis in Post-Tuberculosis Patients: A Retrospective Observational Study

Ming Wu[1], Yan Na Yang[1], Fei Wang[1], Ju Rong Yan[1], Rui Yang[2], ChengQing Yang[3], Yi Ren[1]

[1]Department of Clinical Laboratory, Wuhan Pulmonary Hospital, Wuhan, Hubei Province, 430030, People's Republic of China; [2]Department of Orthopedics, People's Hospital of Dongxihu District, Wuhan, Hubei Province, 430040, People's Republic of China; [3]Respiratory Ward Two, Wuhan Pulmonary Hospital, Wuhan, Hubei Province, 430030, People's Republic of China

Correspondence: ChengQing Yang, Respiratory Ward Two, Wuhan Pulmonary Hospital, Wuhan, Hubei Province, 430030, People's Republic of China, Email clarify719@163.com; Yi Ren, Department of Clinical Laboratory, Wuhan Pulmonary Hospital, Wuhan, Hubei Province, 430030, People's Republic of China, Tel +86-13807196283, Email menease@sina.com

International Journal of General Medicine downloaded from https://www.dovepress.com/
For personal use only.

Objective: The recurrence rate of post-tuberculosis chronic pulmonary aspergillosis (post-TB CPA) is alarmingly high. This study aims to establish a risk prediction model utilizing machine learning algorithms to forecast the one-year recurrence risk of post-TB CPA.

Methods: This retrospective study included all patients diagnosed with pulmonary tuberculosis complicated by chronic pulmonary aspergillosis at Wuhan Pulmonary Hospital in 2022. Ultimately, 220 patients were included for the significance analysis.The Least Absolute Shrinkage and Selection Operator LASSO regression analysis was utilized to select 8 variables associated with the recurrence of tuberculosis complicated by chronic pulmonary aspergillosis. Four machine learning algorithms were compared to predict the recurrence risk in patients with this complication, with their performance evaluated using the receiver operating characteristic curve, area under the curve (AUC), calibration curve analysis, and decision curve analysis.

Results: LASSO regression analysis identified chronic obstructive pulmonary disease (COPD), chronic fibrotic pulmonary aspergillosis (CFPA), progressive pleural hypertrophy, fungal culture results, age, disease duration, emphysema and treatment duration as factors related to the recurrence risk of tuberculosis complicated by chronic pulmonary aspergillosis. The logistic regression model demonstrated the best performance, it outperformed the other three models by achieving the highest AUC of 0.779 on the internal validation set and 0.819 in the test cohort. The calibration curve indicated a strong correlation between the actual and predicted probabilities, while the decision curve analysis revealed significant clinical benefits.

Discussion: In this study, we developed a disease recurrence prediction model using machine learning techniques. This model aims to assist clinicians in identifying the most relevant risk factors associated with the recurrence of tuberculosis complicated by chronic pulmonary aspergillus. It facilitates the formulation of targeted and effective re-examination plans for discharged patients, ultimately reducing the recurrence rate after discharge and enhancing the quality of life for these patients.

Keywords: post-TB CPA, machine learning, predictive model, risk factors, recurrence risk, AUC

Introduction

Tuberculosis (TB) is a significant public health concern. It is estimated that in 2025, the number of new TB cases in China will reach Nearly 800,000, with a death toll of 30,000.[1] Pulmonary tuberculosis (PTB) is caused by the invasion of *Mycobacterium tuberculosis* into the lungs. It is the most prevalent form of TB, accounting for over 80% of all cases.[2,3] More than two-thirds of PTB patients continue to experience extensive lung structural damage following treatment.[4,5] Post-tuberculosis lung disease (PTLD) refers to a group of lung diseases characterized by chronic respiratory system abnormalities, either partially or entirely resulting from PTB, with or without clinical

Received: 2 June 2025
Accepted: 21 November 2025
Published: 3 December 2025

© 2025 Wu et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

symptoms.[6] The lung structural damage caused by tuberculosis can be further complicated by infections such as chronic pulmonary aspergillosis (CPA).[7] As a complication of PTB, CPA affects approximately 3 million individuals worldwide, with a global prevalence rate of 42 per 100,000.[8,9] Research indicates that the 5-year mortality rate for CPA ranges from 50% to 85%,[10] contributing to its status as a global burden. This complication is referred to as post-tuberculosis chronic pulmonary aspergillosis (post-TB CPA). Notably, China continues to bear a high burden of tuberculosis on a global scale. Post-TB CPA patients are not uncommon; however, there is a notable lack of comprehensive literature on this topic. The coexistence of these two conditions complicates both diagnosis and treatment. Currently, there are no established diagnostic and treatment guidelines for post-TB CPA in China. In recent years, as the survival rate of tuberculosis patients has improved, the incidence of CPA among those with pulmonary tuberculosis complications has also significantly increased.[11,12]

Factors influencing the development of chronic pulmonary aspergillosis in patients with pulmonary tuberculosis may include the patient's immune status, history of pulmonary diseases, environmental exposures, and genetic susceptibility.[13,14] These factors are complex and severely impact the quality of life of affected individuals. Preventing recurrence is a key objective in the management of chronic pulmonary aspergillosis. Recurrence can lead to significant distress for both patients and their families.[15] This study aims to analyze the clinical characteristics of CPA patients following tuberculosis, identify the most relevant factors associated with recurrence, and ultimately enhance the quality of life for discharged patients.

Traditional research methods exhibit certain limitations in identifying the interactions among complex factors; however, the application of machine learning technology has introduced new opportunities in this field. Machine learning is adept at handling large-scale and multi-dimensional data.[16,17] By establishing intricate models to identify potential influencing factors and their interrelationships, it can enhance the accuracy of disease prediction and early diagnosis and also identify potential recurrence factors of the disease.[17–19] In recent years, the use of machine learning in medical research has gained significant traction. Through the construction and optimization of machine learning models, researchers can effectively identify recurrence factors related to pulmonary tuberculosis complicated by chronic pulmonary aspergillosis, thereby supporting clinical decision-making.[20] This study retrospectively analyzed the clinical characteristics and treatment data of 220 patients with pulmonary tuberculosis complicated by chronic pulmonary aspergillosis who were admitted to Wuhan Pulmonary Hospital in 2022. Machine learning was employed to analyze the factors influencing the recurrence of post-TB CPA patients within one year, aiming to enhance clinical management and post-discharge follow-up for these patients. This review provides valuable insights.

Materials and Methods

Study Population

A retrospective study method was employed in this research. From January to December 2022, the inpatient electronic medical record system of Wuhan Pulmonary Hospital was utilized to identify patients diagnosed with pulmonary fungal infections, specifically Aspergillus pneumonia, invasive pulmonary aspergillosis, allergic bronchopulmonary aspergillosis, and disseminated aspergillosis. This selection encompassed patients with potential diagnoses of pulmonary aspergillosis. After applying the weight removal process, the diagnoses of all patients were re-evaluated based on predefined inclusion and exclusion criteria. The inclusion criteria were as follows:1.Chest CT scans must exhibit imaging characteristics consistent with chronic pulmonary aspergillosis (CPA); 2. There must be direct or immunological evidence of *Aspergillus* infection, meeting at least one of the following criteria: ① Positive *Aspergillus* culture from sputum or bronchoalveolar lavage fluid (BALF); ② Detection of *Aspergillus* nucleic acid in sputum or BALF; ③Positive galactomannan antigen test (GM) in BALF; ④ Presence of *Aspergillus*-specific IgG antibodies in serum; ⑤ Pathological confirmation of *Aspergillus* infection in lung tissue; 3. The disease duration must be at least three months. The exclusion criteria included: 1. Invasive pulmonary aspergillosis;2.Allergic bronchopulmonary aspergillosis;3. Absence of *Aspergillus* infection; 4.Suspected *Aspergillus* infections that remain undiagnosed.[21,22]

Data Collection

The clinical data of 220 patients diagnosed with pulmonary tuberculosis (PTB) combined with CPA were collected. This dataset includes various parameters such as gender, age, previous medical history, presenting symptoms, disease duration,

imaging findings, laboratory test classifications, and follow-up outcomes. The follow-up assessments focused on patient stability and recurrence of the disease. Informed consent was obtained from all participants, and the study adhered to the ethical guidelines established by Wuhan Pulmonary Hospital.

Model Construction

The Beckman Coulter Dx AI platform (https://www.xsmartanalysis.com/beckman/login/) was utilized for statistical analysis. Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis was employed to identify factors associated with the recurrence of post-TB CPA. This study implemented four machine learning classification algorithms, namely XGBoost, Random Forest, logistic regression, and Support Vector Machine (SVM), to construct the machine learning model. The training dataset was used for optimal model selection, while the test dataset was utilized to validate the best-performing model. Evaluation metrics, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and area under the receiver operating characteristic curve (AUC), were employed to compare the efficacy of the models.

Statistical Analysis

Continuous variables were standardized, while categorical variables underwent exclusive encoding. Continuous variables were represented using the median (M) and interquartile range (IQR), whereas categorical variables were expressed as counts and percentages. The differences between groups for continuous variables were analyzed using either Student's *t*-test or the Mann–Whitney test, while categorical variables were examined using Fisher's exact test or the chi-squared test.

Key variables for constructing the prediction model were selected using Lasso regression. The optimal algorithm was then identified by comparing multiple modeling algorithms and was used to build the prediction model individually. Discriminative ability was quantified by the area under the receiver operating characteristic curve (AUC).

All statistical analyses were conducted using Python version 3.11.4, R version 4.2.3 and glmnet version 4.1.8. Results were deemed statistically significant at a threshold of $P < 0.05$.

Result

Basic Characteristics of the Patients

A total of 220 patients were enrolled in this study, comprising 141 males (64%) and 79 females (36%). The median age of the patients was 60 years (Table S1). Continuous variables were presented as mean±standard deviation (SD) for normally distributed variables or median with interquartile range (OR) for non-normally distributed variables, and categorical variables were presented as percentage frequencies. Demographic and laboratory tests were compared using a t-testor Mann–Whitney *U*-test for continuous and chi-square test for categorical variables. We applied Shapiro–Wilk Nornality test to test for normality, and Levene's-test to test homogeneity of variances.Then, nomally distributed variables were compared by the Student's *t*-test, and non-normally distributed variables were compared by MannWhitney *U*-test. In the significant analysis and comparison, it was observed that the combination of bronchiectasis, Non-tuberculous mycobacteria (NTM), and chronic obstructive pulmonary disease (COPD), along with chronic fibrotic pulmonary aspergillosis (CFPA), showed P values for CFPA, number of cavities, progressive pleural hypertrophy, emphysema, fungal culture, age, disease duration, Erythrocyte Sedimentation Rate(ESR), Serum Fungal Galactomannan Test (Serum GM), and treatment course all less than 0.05 in Table 1, indicating statistical significance.

Feature Screening

To identify the most relevant predictors for recurrence and to avoid overfitting, feature selection was performed using the Least Absolute Shrinkage and Selection Operator (Lasso) regression with a binomial deviance penalty. The optimal regularization parameter (λ) was determined via 10-fold cross-validation.The cross-validation results are presented in Figure 1A, which plots the binomial deviance against Log(λ). The left vertical dashed line indicates the value of λ (λ_min = 0.035) that achieved the minimum mean cross-validated error. The right vertical dashed line corresponds to the value of λ (λ_1se = 0.067), which represents the most parsimonious model within one standard error of the minimum error.The coefficient profiles for all

Table 1 The Results of the Significance Analysis

Variable	Overview (n=220)	Follow-up 0 (n=134)	Follow-up 1 (n=86)	Statistics	p	Method
Sex,n(%)	141(64.091)	78(58.209)	63(73.256)	5.153	0.023	Chi-square test
Bronchiectasis,n(%)	189(85.909)	121(90.299)	68(79.070)	5.456	0.02	Chi-square test
NTM,n(%)	201(91.364)	127(94.776)	74(86.047)	5.059	0.024	Chi-square test
COPD,n(%)	171(77.727)	119(88.806)	52(60.465)	24.303	<0.001	Chi-square test
Classified aspergillus nodule,n(%)	195(88.636)	110(82.090)	85(98.837)	14.587	<0.001	Chi-square test
Solitary aspergilloma,n(%)	189(85.909)	110(82.090)	79(91.860)	4.131	0.042	Chi-square test
CFPA,n(%)	181(82.273)	121(90.299)	60(69.767)	15.139	<0.001	Chi-square test
Number of cavities,n(%)	27(12.273)	26(19.403)	1(1.163)	16.187	<0.001	Chi-square test
Air crescent sign,n(%)	73(33.182)	54(40.299)	19(22.093)	7.831	0.005	Chi-square test
Intracavitary septum/septation,n(%)	54(24.545)	45(33.582)	9(10.465)	15.114	<0.001	Chi-square test
Progressive pleural thickening,n(%)	54(24.545)	49(36.567)	5(5.814)	26.749	<0.001	Chi-square test
Aspergillus nodule,n(%)	172(78.182)	98(73.134)	74(86.047)	5.12	0.024	Chi-square test
Emphysema,n(%)	99(45.000)	76(56.716)	23(26.744)	19.013	<0.001	Chi-square test
Fibrosis,n(%)	91(41.364)	66(49.254)	25(29.070)	8.798	0.003	Chi-square test
Fungal culture,n(%)	166(75.455)	109(81.343)	57(66.279)	6.418	0.011	Chi-square test
Age,median[IQR]	60.000[52.000,66.000]	57.000[43.000,64.000]	64.000[57.000,68.000]	−4.276	<0.001	Mannwhitney-U
Disease duration (months),median[IQR]	12.000[3.000,36.000]	10.000[3.000,24.000]	15.000[6.000,48.000]	−2.931	0.003	Mannwhitney-U
N,median[IQR]	4.260[3.040,5.640]	4.110[2.800,5.330]	4.670[3.470,6.700]	−2.502	0.012	Mannwhitney-U
L,median[IQR]	1.220[0.890,1.650]	1.270[0.960,1.700]	1.020[0.830,1.470]	2.844	0.004	Mannwhitney-U
Hb,mean(±SD)	116.609±20.797	119.993±18.284	111.337±23.236	2.907	0.004	Welch's *t*-test
Alb,mean(±SD)	34.980±5.246	35.781±5.160	33.733±5.134	2.866	0.005	*t*-test
ESR,median[IQR]	35.000[14.000,66.000]	27.000[12.000,58.000]	50.000[20.000,84.000]	−3.645	<0.001	Mannwhitney-U
hCRP,median[IQR]	13.470[2.000,38.680]	6.500[1.350,29.970]	20.650[5.240,49.530]	−3.574	<0.001	Mannwhitney-U
Serum GM,median[IQR]	0.090[0.030,0.160]	0.070[0.030,0.140]	0.110[0.050,0.260]	−3.219	0.001	Mannwhitney-U
BFGm,median[IQR]	1.020[0.310,2.870]	1.270[0.350,3.420]	0.760[0.230,2.550]	1.984	0.047	Mannwhitney-U
Treatment course,median[IQR]	3.000[0.000,6.000]	6.000[2.000,6.000]	1.000[0.000,6.000]	4.067	<0.001	Mannwhitney-U
Pulmonary nodule,n(%)	220(100.000)	134(100.000)	86(100.000)	0	1	Chi-square test
Bronchial tuberculosis,n(%)	164(74.545)	97(72.388)	67(77.907)	0.841	0.359	Chi-square test
Pulmonary tumor,n(%)	218(99.091)	134(100.000)	84(97.674)	nan	nan	Chi-square test
Diabetes mellitus,n(%)	176(80.000)	103(76.866)	73(84.884)	2.105	0.147	Chi-square test
Pneumoconiosis,n(%)	211(95.909)	130(97.015)	81(94.186)	1.068	0.301	Chi-square test
Pulmonary sarcoidosis,n(%)	220(100.000)	134(100.000)	86(100.000)	0	1	Chi-square test
Interstitial pneumonia,n(%)	216(98.182)	132(98.507)	84(97.674)	0.204	0.652	Chi-square test
Fever,n(%)	198(90.000)	119(88.806)	79(91.860)	0.543	0.461	Chi-square test
CCPA,n(%)	100(45.455)	64(47.761)	36(41.860)	0.736	0.391	Chi-square test
SAIA,n(%)	215(97.727)	131(97.761)	84(97.674)	0.002	0.966	Chi-square test
Right lung (imaging findings),n(%)	58(26.364)	36(26.866)	22(25.581)	0.045	0.833	Chi-square test
Left lung,n(%)	88(40.000)	65(48.507)	23(26.744)	10.338	0.001	Chi-square test
Lung field,n(%)	2(0.909)	2(1.493)	0(0.000)	nan	nan	Chi-square test
Aspergillus IgG,n(%)	70(31.818)	45(33.582)	25(29.070)	0.492	0.483	Chi-square test
Aspergillus nucleic acid,n(%)	101(45.909)	56(41.791)	45(52.326)	2.341	0.126	Chi-square test
WBC,median[IQR]	6.230[4.930,7.810]	6.180[4.580,7.650]	6.230[5.280,8.210]	−1.849	0.065	Mannwhitney-U
PLT,median[IQR]	224.000[172.000,286.000]	223.000[174.000,285.000]	219.000[170.000,288.000]	0.226	0.822	Mannwhitney-U
Aspergillus IgG level,median[IQR]	1.115[0.604,1.474]	1.006[0.604,1.450]	1.191[0.600,1.563]	−1.204	0.229	Mannwhitney-U

candidate variables across the range of λ values are illustrated in Figure 1B. Each path represents the trajectory of a variable's coefficient as the penalty increases (moving rightward on the Log(λ) axis). The two vertical lines from the cross-validation plot are superimposed, showing the exact set of variables retained at each critical λ value. Variables whose coefficients shrink to zero are effectively excluded from the model.At λ_min = 0.035, the Lasso model selected 13 variables with non-zero coefficients. To enhance model simplicity and generalizability, the model based on λ_1se = 0.067 was adopted for further analysis. At this penalty level, the coefficients of eight variables remained non-zero, identifying them as key predictors. As detailed in Table 2 (under the "SEM of minimal distance" column), these factors and their coefficients are: chronic obstructive pulmonary disease (COPD, 0.610), chronic pulmonary aspergillosis (CFPA, 0.458), progressive pleural thickening (0.778), emphysema (0.045), fungal culture (0.089), age (0.010), disease duration (0.001), and treatment course (−0.082). The

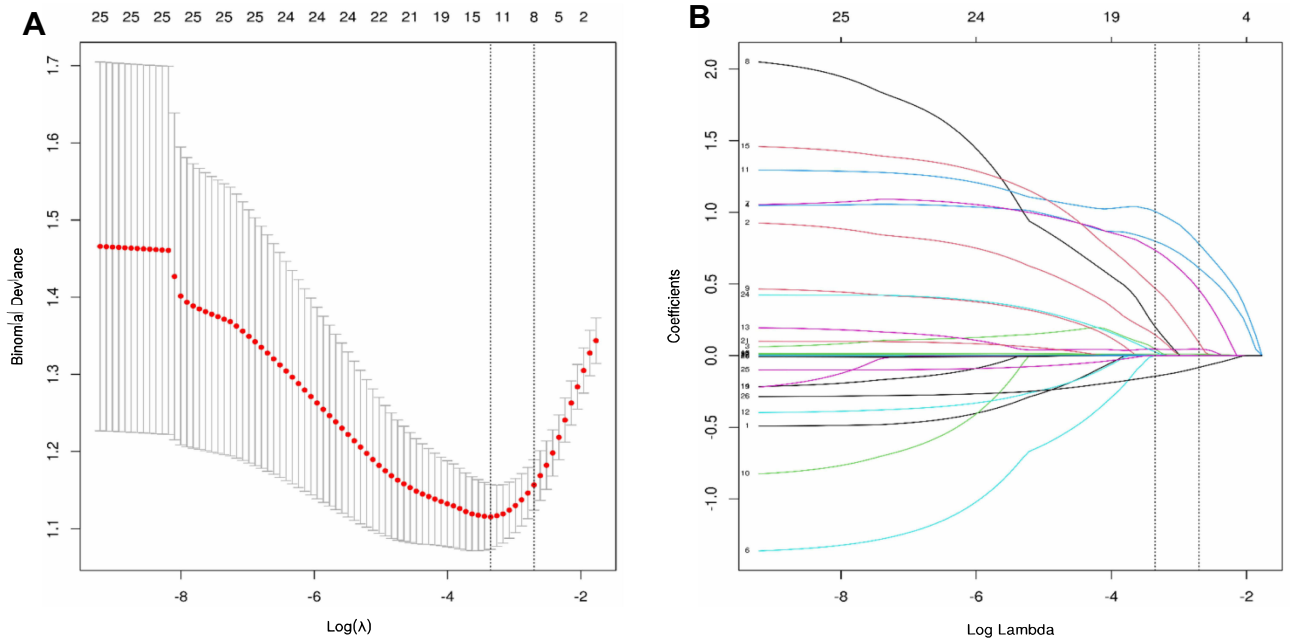

Figure 1 Least absolute shrinkage and selection operator (LASSO) regression analysis and 5-fold cross-validation for selecting factors associated with the recurrent patients of post-TB CPA patients.(**A**) A joint plot was created based on the log-likelihood.The minimum standard is on the left line and the 1-SE standard is on the right line. In the current study, we selected 7 non-zero predictors according to the 1-SE standard. SE, the standard error.(**B**) Bias selection of the tuning parameter (lambda) in LASSO regression based on the minimum standard (left dashed line) and 1-SE (standard error) standard (right dashed line).

magnitude of these coefficients reflects the strength of association with recurrence, while the sign indicates the direction (positive for risk factors, negative for protective factors). The negative coefficient for "treatment course" suggests a protective effect against recurrence.

Table 2 The Coefficients of Lasso Regression

Name	LMS Coefficient	SEM of Minimal Distance
(Intercept)	−5.75	−3.609
Bronchiectasis	0.144	0
NTM	0.046	0
COPD	0.798	0.61
Typed aspergilloma	0	0
CFPA	0.735	0.458
Number of cavities	0.201	0
Air crescent sign	0	0
Intracavitary septum	0	0
Progressive pleural thickening	1.006	0.778
Aspergillus nodule	0	0
Emphysema	0.046	0.045
Fibrosis	0	0
Fungal culture	0.47	0.089
Age	0.011	0.01
Disease duration	0.003	0.001
N	0	0
L	0	0

(*Continued*)

Table 2 (Continued).

Name	LMS Coefficient	SEM of Minimal Distance
Hb	0	0
Alb	0	0
ESR	0.002	0
hCRP	0	0
Serum GM	0.026	0
BFGm	0	0
Treatment course	−0.144	−0.082

Identification of the Optimal Model

Four machine learning models—namely XGBoost, Random Forest, Logistic Regression, and Support Vector Machine (SVM)—were employed to analyze and compare the evaluation indicators among the models in order to identify the optimal predictive model. To mitigate the risk of overfitting and to select the best model, a 5-fold cross-validation was performed using the training dataset. Figure 2 illustrates the performance of the four machine learning models in both the training and validation cohorts. Subsequently, the average values of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and area under the curve (AUC) for the four machine learning models were calculated. The performance metrics of the logistic model in the validation set, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and AUC (95% CI), are presented in Table 3. Among the four comparative models, logistic regression ranked third in AUC on the training set, yet its training performance remained notably high with an AUC of 0.819. More importantly, logistic regression outperformed the other three models by achieving the highest AUC of 0.779 on the internal validation set. As the internal validation set is widely recognized as a far more critical indicator for evaluating a model's generalization potential,since it simulates the model's performance on unseen data more accurately than the training set and its stronger robustness and adaptability to non-training data. Thus, we rigorously confirm that logistic regression is the optimal model for this study.

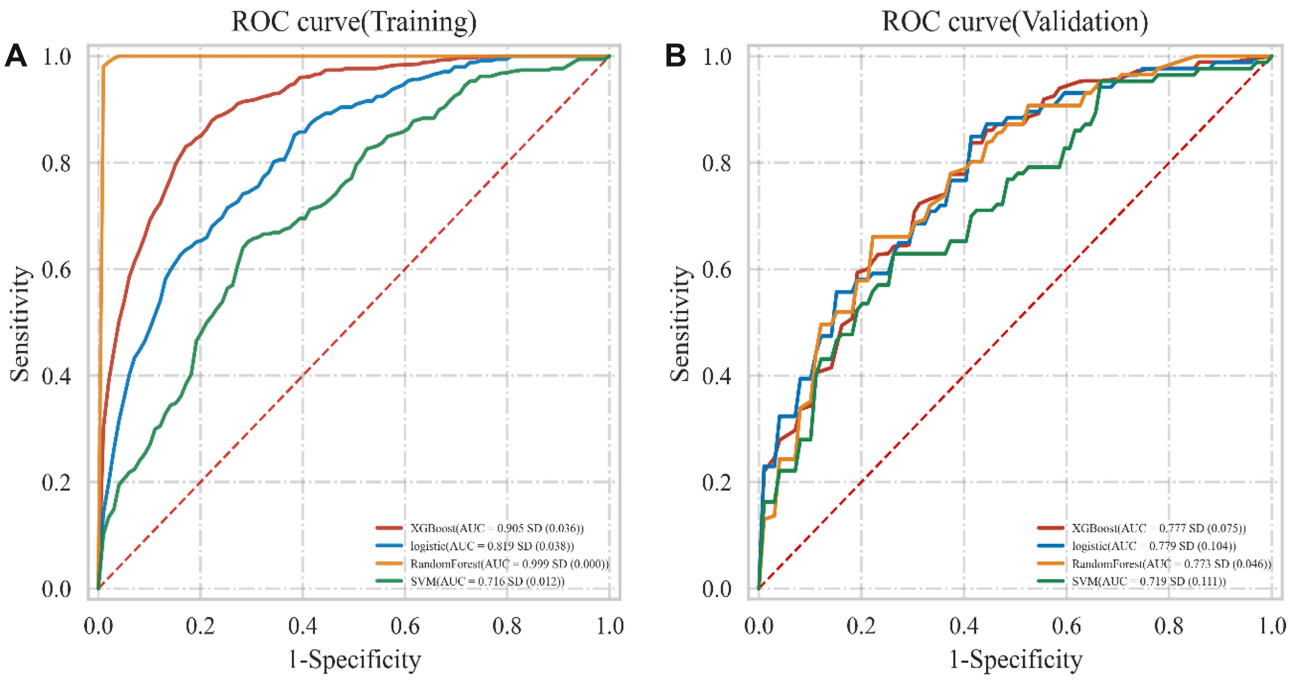

Figure 2 Multiple model comparison results. (**A**). ROC curve comparison of training set in multiple machine algorithms. (**B**). ROC curve comparison of validation set in multiple machine algorithms.

Table 3 Comparison of Multiple Machine Learning Evaluation Indexes Between Training Set and Validation Set

	Model	AUC(SD)	Cutoff(SD)	Accuracy Rating(SD)	Sensitivity (SD)	Specificity (SD)	Positive Predictive Value(SD)	Negative Predictive Value(SD)	F1 score (SD)	Kappa (SD)
Training set	XGBoost	0.905(0.036)	0.412(0.039)	0.825(0.051)	0.893(0.030)	0.782(0.094)	0.734(0.075)	0.919(0.016)	0.802(0.041)	0.650(0.090)
	logistic	0.819(0.038)	0.405(0.062)	0.749(0.046)	0.777(0.095)	0.732(0.113)	0.664(0.075)	0.843(0.045)	0.708(0.041)	0.493(0.080)
	RF	0.999(0.000)	0.444(0.042)	0.986(0.008)	1.000(0.000)	0.978(0.013)	0.967(0.018)	1.000(0.000)	0.983(0.010)	0.972(0.016)
	SVM	0.716(0.012)	0.394(0.009)	0.698(0.006)	0.643(0.034)	0.733(0.025)	0.608(0.013)	0.762(0.012)	0.624(0.012)	0.372(0.011)
Validation set	XGBoost	0.777(0.075)	0.412(0.039)	0.682(0.069)	0.697(0.127)	0.672(0.119)	0.583(0.080)	0.781(0.055)	0.629(0.084)	0.357(0.132)
	logistic	0.779(0.104)	0.405(0.062)	0.686(0.096)	0.731(0.143)	0.656(0.103)	0.580(0.099)	0.796(0.092)	0.644(0.113)	0.370(0.192)
	RF	0.773(0.046)	0.444(0.042)	0.700(0.068)	0.627(0.153)	0.747(0.091)	0.617(0.104)	0.767(0.078)	0.614(0.104)	0.371(0.151)
	SVM	0.719(0.111)	0.394(0.009)	0.682(0.066)	0.605(0.052)	0.732(0.108)	0.606(0.102)	0.741(0.035)	0.601(0.060)	0.338(0.121)

https://doi.org/10.2147/IJGM.S544265

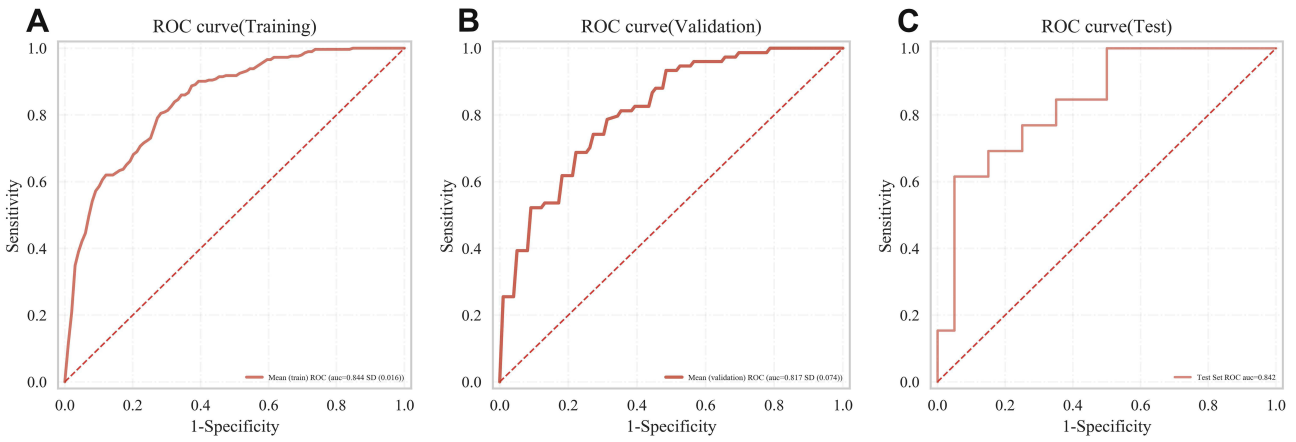

Figure 3 Performance of the prediction model (**A**) Receiver operating characteristic (ROC) curve of the training cohort; (**B**) ROC curve of the validation cohort; (**C**) ROC curve of the testing cohort.

Analysis and Assessment of the Logistic Model

We employ the logistic model as the final classification method for binary classification.To ensure the objectivity and reliability of model performance evaluation, this study adopted a dual validation strategy combining "independent test set + cross-validation. Specifically, 15% of the total data was randomly extracted as the Independent Test Set; this part of the data was not involved in model training throughout the process and was only used for External Validation under simulated real-world scenarios to verify the model's generalization ability.The remaining 85% of the dataset was used for model construction and internal evaluation, implemented through 5-Fold Cross-Validation: this part of the data was evenly divided into 5 subsets. In each iteration, 4 subsets served as the training set and 1 subset as the internal validation set. This process was repeated 5 times to achieve full sample coverage. Finally, the mean and standard deviation of the 5 validation results were used to comprehensively evaluate the model's stability and internal consistency. Calibration curvesd for each set in Figure 3A–C. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and AUC (95% CI) of the model in the test, training, and validation sets are presented in Table 4. The AUC of the test set (0.842), which is very close to the training set's 0.844. Notably, the test set (0.842) was significantly higher than the validation set (0.817). These results collectively demonstrate that our model exhibits excellent performance—it not only avoids overfitting but also achieves superior generalization ability in real-world-like scenarios compared to internal validation, fully confirming the model's reliability and effectiveness.

The SHAP Summary Graph

As illustrated in Figure 4A and B, the SHAP algorithm was used to interpret feature importance in the logistiregression model. COPD, CFPA, progressive pleural thickening, emphysema, fungal culture, age, disease duration, and treatment course were identified as the most influential features fopredicting recurrence of chronic pulmonary aspergillosis in post-tuberculosis patients.A larger mean absolute Shapley value indicated a greater impact on the model'predictions.To clarify, these figures decode the risk factors for recurrent post-tuberculosis CPA via SHAP: Figure 4A (SHAP summary scatter plot):Displays how features influence recurrence risk—red/blue colors denote high/low feature values, and SHAP

Table 4 Diagnostic Efficacy of the Logistic Model in the Testing and Validation Cohorts

	AUC	Cutoff	Accuracy Rating	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	F1 score
Training	0.862	0.374	0.789	0.817	0.771	0.692	0.876	0.746
Validation	0.842	0.374	0.732	0.760	0.714	0.630	0.832	0.683
Testing	0.715	0.398	0.576	0.667	0.5	0.526	0.643	0.588

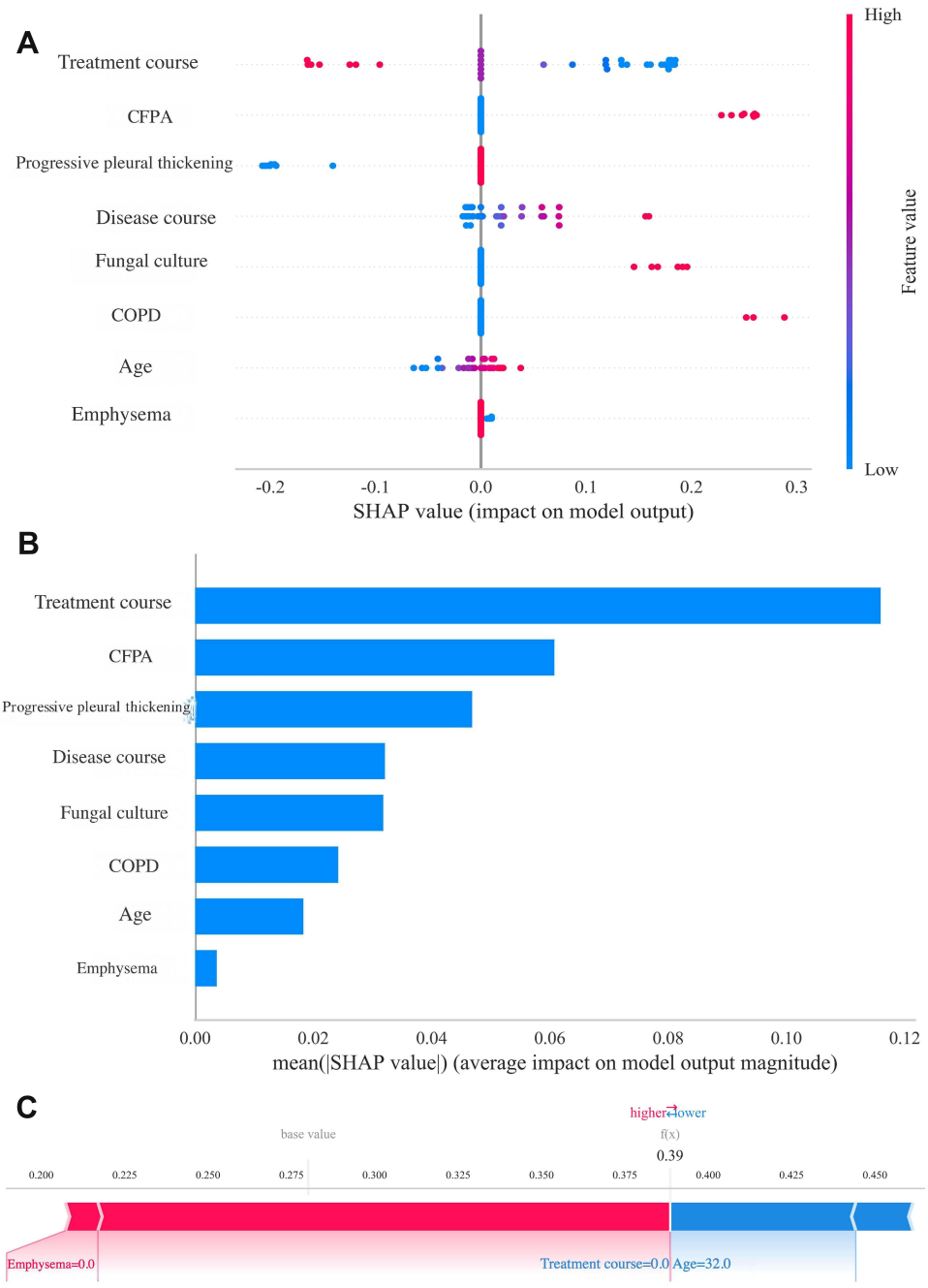

Figure 4 Summary plots of visualize SHAP values. In Figure (**A**) the position of the point along the X-axis represents the actual SHAP value, indicating the influence of a specific feature on the model output of that specific patient. Mathematically speaking, this corresponds to the logarithm of the recurrence risk among patients, which means that a higher SHAP value indicates a higher recurrence risk compared to patients with a lower SHAP value. Features are organized along the Y-axis according to their importance and are determined by the average of their absolute Shapley values. The higher the position of the feature in the graph is, the more significant the influence on the model will be. Figure (**B**) shows the SHAP importance graph of the logistic algorithm. (**C**) Contribution level of features in logistic model.

values show their push/pull on risk. Figure 4B (SHAP feature importance bar plot): Ranks features by their average impact on risk, with Treatment course being the most critical and Emphysema the least. Figure 4C (SHAP force plot): Illustrates how features collectively determine an individual's recurrence risk (f(x)=0.39). Together, they comprehensively unravel the model's reasoning for predicting CPA recurrence risk.

https://doi.org/10.2147/IJGM.S544265

Discussion

Chronic Pulmonary Aspergillosis (CPA), as a complication of pulmonary tuberculosis (PTB), imposes a considerable burden worldwide.[20] Epidemiological data indicate that 5% to 35% of patients develop CPA as a consequence of tuberculosis, with half of the global CPA burden attributable to this disease.[13,23] Although there have been several studies on CPA following tuberculosis in international contexts, relevant research reports from China are scarce. Patients suffering from post-TB CPA frequently exhibit symptoms such as chronic cough, expectoration, and hemoptysis. Additionally, some patients may present with tuberculosis-related systemic symptoms, including fever, night sweats, and weight loss,[24] though these clinical manifestations tend to be more severe. The clinical presentation of tuberculosis combined with CPA poses a complex clinical challenge.[24,25] In addition to diagnosis and treatment, preventing recurrence is also of paramount importance.

In recent years, the widespread application of machine learning technology in the medical field has resulted in a growing number of studies aimed at utilizing machine learning models to predict the likelihood of diseases and clinical outcomes in patients with TB and CPA.[13,26] Research indicates that machine learning models, which incorporate clinical features and biomarkers, hold significant promise for predicting disease risk and prognosis in these patient populations.[27,28] For instance, Yan et al employed logistic regression to estimate the probability of tuberculosis in conjunction with CPA, taking into account the patient's disease history, lung cavitation, and surgical background. Their model achieved an AUC of 0.86.[28] Additionally, Ren et al utilized serum cytokine biomarkers, including IL-6 and IL-8, as diagnostic criteria for pulmonary tuberculosis and chronic pulmonary aspergillosis.[29]

This study retrospectively analyzed the clinical data of 72 patients with tuberculosis complicated by CPA at Wuhan Pulmonary Hospital in 2022, including various laboratory examination indicators. A one-year recurrence risk prediction model for the disease was constructed using machine learning algorithms. Four different machine learning models were compared to identify the most relevant risk factors for recurrence in patients with tuberculosis complicated by CPA. Seven significant influencing factors were identified: COPD, CPA, progressive pleural hypertrophy, fungal culture, age, disease duration, and treatment duration. By analyzing SHAP values, we quantified the contribution of each feature to the model output, thus determining the most relevant predictor.[30] This analytical method provides a deeper understanding of how each factor influences the predictive results of the model. A comprehensive evaluation of these factors is crucial for clinical decision-making and patient management.

The results of the SHAP analysis indicate that the treatment course is the most significant factor. A full treatment course can reduce the risk of recurrence, which aligns with previous studies.[22] The duration of the treatment course directly impacts the treatment outcome and the risk of recurrence. An appropriate treatment duration is essential for controlling the condition and preventing recurrence. Progressive pleural hypertrophy is the second most relevant factor. Studies have demonstrated that the absence of cavitation on chest X-rays and the presence of 100% pleural thickening serve as negative predictive values for CPA.[31] Therefore, this study recommends that discharged patients undergo chest X-ray examinations every three months, with particular attention to pleural hypertrophy, as this can help prevent recurrence.COPD and chronic fibrotic pulmonary aspergillus disease are the third risk factors for recurrence and warrant significant attention. Pneumothorax is a common complication of post-tuberculosis CPA.Tuberculosis-induced lung damage (cavities, fibrosis) plus Aspergillus-induced inflammation weakens lung tissue, easily causing rupture and pneumothorax.[22] Conversely, pneumothorax worsens PT-CPA patients' respiratory dysfunction and indicates poor disease control, with high recurrence risk.It is advised that such patients complete the full course of treatment during their hospital stay, maintain good compliance, and increase the frequency of follow-ups and regular reexaminations after discharge. Patients with a longer disease duration and those with positive fungal cultures face an elevated risk of recurrence and should use medications judiciously. Age is also a significant factor; it is one of the critical determinants influencing the development of chronic pulmonary aspergillosis, as confirmed by numerous studies.[32] As individuals age, the immune system's functionality declines, rendering elderly patients more susceptible to infections.

While this research has yielded significant results, it is not without limitations. The study was conducted at a single center, which may limit the generalizability of the model to other regions or populations. Additionally, the absence of long-term follow-up precludes an evaluation of the model's predictive capabilities regarding long-term prognosis.

However, this study also presents notable strengths: the application of multiple machine learning algorithms for model construction and comparison enhances the stability and reliability of the models. By employing 5-fold cross-validation and a comprehensive evaluation using multiple indices, the optimal logistic regression model was identified, effectively mitigating issues of overfitting and model selection bias.[33] The model demonstrates commendable predictive performance on both the training and validation sets, suggesting a degree of generalization ability.[34] The contribution of each feature to the model was analyzed using SHAP values, providing valuable insights for clinical decision-making. Despite the identified limitations, through rigorous methodological design and multi-faceted model evaluation, a relatively stable and reliable one-year recurrence risk prediction model for pulmonary tuberculosis combined with chronic pulmonary aspergillosis has been established. Future research should aim to expand the sample size, conduct multi-center validations, and investigate the predictive effects and long-term prognoses of this model across diverse populations.

Conclusion

This study constructed a machine learning-based one-year recurrence prediction model for post-TB CPA using 220 retrospective cases from Wuhan Pulmonary Hospital; LASSO regression identified eight key risk factors (COPD, CFPA, progressive pleural hypertrophy, etc)., and among four algorithms, the logistic regression model performed best (AUC: 0.779 internal validation, 0.819 test cohort) with good calibration and clinical benefits. In conclusion, this model effectively identifies high-risk patients and quantifies recurrence risk, providing a practical tool for personalized follow-up to reduce recurrence and improve prognosis, while future multi-center prospective studies are needed to validate its generalizability.

Data Sharing Statement

All data supporting the results of this study are available upon request from the corresponding author.

Ethics Statement

This study was approved by the Ethics Committee of [Wuhan Pulmonary Hospital] (Approval No.[Wuhan Ethics (2022) No. 7]). The Ethics Committee waived the requirement for informed consent because of the retrospective nature of the study and the anonymization of all participant data.

Funding

This work was financially supported by the Wuhan Science and Technology Innovation Bureau Project (No. 2024040801020381). Wuhan Municipal Health and Family Planning Commission Scientific research project (No: WX20D75).

Disclosure

The authors declare that they have no competing interests.

References

1. Trajman A, Campbell JR, Kunor T, et al. Tuberculosis. *Lancet*. 2025;405(10481):850–866. doi:10.1016/S0140-6736(24)02479-6
2. Chihota V, Gombe M, Gupta A, et al. Tuberculosis preventive treatment in high TB-burden settings: a state-of-the-art review. *Drugs*. 2025;85(2):127–147. doi:10.1007/s40265-024-02131-3
3. Tayal A, Kabra SK. Tuberculosis preventive treatment. *Indian J Pediatr*. 2023;91(8):823–829. doi:10.1007/s12098-023-04969-z
4. Dheda K, Mirzayev F, Cirillo DM, et al. Multidrug-resistant tuberculosis. *Nat Rev Dis Primers*. 2024;10(1):22. doi:10.1038/s41572-024-00504-2
5. Janssen S, Murphy M, Upton C, Allwood B, Diacon AH. Tuberculosis: an update for the clinician. *Respirology*. 2025;30(3):196–205. doi:10.1111/resp.14887
6. Bongomin F, Denning DW. Chronic pulmonary aspergillosis: a neglected post-TB lung disease. *Int J Tuberc Lung Dis*. 2024;28(6):314–315. doi:10.5588/ijtld.24.0165
7. Denning DW. Global incidence and mortality of severe fungal disease. *Lancet Infect Dis*. 2024;24(7):e428–e438. doi:10.1016/S1473-3099(23)00692-8
8. Wichmann D, Hoenigl M, Koehler P, et al. Diagnosis and treatment of invasive pulmonary aspergillosis in critically ill intensive care patients: executive summary of the German national guideline (AWMF 113-005). *Infection*. 2025;53(4):1299–1310. doi:10.1007/s15010-025-02572-2
9. Ledford DK, Kim T-B, Ortega VE, Cardet JC. Asthma and respiratory comorbidities. *J Allergy Clin Immunol*. 2024;155(2):316–326. doi:10.1016/j.jaci.2024.11.006

10. Sengupta A, Ray A, Upadhyay AD, et al. Mortality in chronic pulmonary aspergillosis: a systematic review and individual patient data meta-analysis. *Lancet Infect Dis*. 2025;25(3):312–324. doi:10.1016/S1473-3099(24)00567-X
11. Rozaliyani A, Setianingrum F, Isbaniah F, et al. A silent threat in post-tuberculosis patients: chronic pulmonary aspergillosis survey in multiple regions of Indonesia (I-CHROME study). *J Fungi*. 2025;11(5). doi:10.3390/jof11050329
12. Jaggi TK, Agarwal R, Tiew PY, et al. Fungal lung disease. *Eur Respir J*. 2024;64(5):2400803. doi:10.1183/13993003.00803-2024
13. Madden A, Ofori SK, Budu M, Sisay E, Dooley B, Murray MB. A systematic review of chronic pulmonary aspergillosis among patients treated for pulmonary tuberculosis. *Clin Infect Dis*. 2025;81(4):e163–e171. doi:10.1093/cid/ciaf150
14. Li Y, Ren X, Wang Q, et al. A predictive model for pulmonary aspergillosis in ICU Patients: a multicenter retrospective cohort study. *Infect Drug Resist*. 2025;18:441–454. doi:10.2147/IDR.S493019
15. Bongomin F. Post-tuberculosis chronic pulmonary aspergillosis: an emerging public health concern. *PLoS Pathog*. 2020;16(8):e1008742. doi:10.1371/journal.ppat.1008742
16. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*. 2022;23(1):40–55. doi:10.1038/s41580-021-00407-0
17. Theodosiou AA, Read RC. Artificial intelligence, machine learning and deep learning: potential resources for the infection clinician. *J Infect*. 2023;87(4):287–294. doi:10.1016/j.jinf.2023.07.006
18. Guo Q-H, Xie F-C, Zhong F-M, et al. Application of interpretable machine learning algorithms to predict distant metastasis in ovarian clear cell carcinoma. *Cancer Med*. 2024;13(7):e7161. doi:10.1002/cam4.7161
19. Gu W, Chen Y, Zhu H, et al. Development and validation of CT-based radiomics deep learning signatures to predict lymph node metastasis in non-functional pancreatic neuroendocrine tumors: a multicohort study. *EClinicalMedicine*. 2023;65:102269. doi:10.1016/j.eclinm.2023.102269
20. Neuböck MJ, Günther G, Barac A, et al. Chronic pulmonary aspergillosis as a considerable complication in post-tuberculosis lung disease. *Semin Respir Crit Care Med*. 2024;45(1):102–113. doi:10.1055/s-0043-1776913
21. Denning DW, Cadranel J, Beigelman-Aubry C, et al. Chronic pulmonary aspergillosis: rationale and clinical guidelines for diagnosis and management. *Eur Respir J*. 2016;47(1):45–68. doi:10.1183/13993003.00583-2015
22. Zhong H, Wang Y, Gu Y, et al. Clinical features, diagnostic test performance, and prognosis in different subtypes of chronic pulmonary aspergillosis. *Front Med Lausanne*. 2022;9:811807. doi:10.3389/fmed.2022.811807
23. Rayens E. Estimating mortality in chronic pulmonary aspergillosis. *Lancet Infect Dis*. 2025;25(3):250–251. doi:10.1016/S1473-3099(24)00659-5
24. Nguyen NTB, Le Ngoc H, Nguyen NV, et al. Chronic pulmonary aspergillosis situation among post tuberculosis patients in vietnam: an observational study. *J Fungi*. 2021;7(7). doi:10.3390/jof7070532
25. Tobin EH, Gilotra TS, Baradhi KM. *Aspergilloma*. Treasure Island (FL): StatPearls Publishing; 2024.
26. Soeroso NN, Siahaan L, Khairunnisa S, et al. The association of chronic pulmonary aspergillosis and chronic pulmonary histoplasmosis with MDR-TB patients in indonesia. *J Fungi*. 2024;10(8). doi:10.3390/jof10080529
27. Vithayathil M, Koku D, Campani C, et al. Machine learning based radiomic models outperform clinical biomarkers in predicting outcomes after immunotherapy for hepatocellular carcinoma. *J Hepatol*. 2025;83(4):959–970. doi:10.1016/j.jhep.2025.04.017
28. Karlsson L, Vogel J, Arvidsson I, et al. Machine learning prediction of tau-PET in Alzheimer's disease using plasma, MRI, and clinical data. *Alzheimers Dement*. 2025;21(2):e14600. doi:10.1002/alz.14600
29. Ren W, Li H, Guo C, et al. Serum cytokine biomarkers for use in diagnosing pulmonary tuberculosis versus chronic pulmonary aspergillosis. *Infect Drug Resist*. 2023;16:2217–2226. doi:10.2147/IDR.S403401
30. Bai Q, Chen H, Gao Z, et al. Advanced prediction of heart failure risk in elderly diabetic and hypertensive patients using nine machine learning models and novel composite indices: insights from NHANES 2003–2016. *Eur J Prev Cardiol*. 2025. doi:10.1093/eurjpc/zwaf081
31. Garg M, Bhatia H, Sehgal I, et al. The conundrum of computed tomography findings in chronic pulmonary aspergillosis: insights from 103 cases. *J Thorac Imaging*. 2025;40(5). doi:10.1097/RTI.0000000000000828
32. Chirumamilla NK, Arora K, Kaur M, et al. Innate and adaptive immune responses in subjects with CPA secondary to post-pulmonary tuberculosis lung abnormalities. *Mycoses*. 2024;67(5):e13746. doi:10.1111/myc.13746
33. Zhang X, Xu Z, Yao Y, et al. Validation status of electronic sphygmomanometers in China: a national survey. *Hypertension*. 2025;82(3):532–541. doi:10.1161/HYPERTENSIONAHA.124.24203
34. Zou X, Gomez ZW, Reddy TE, Allen AS, Majoros WH. Bayesian estimation of allele-specific expression in the presence of phasing uncertainty. *Bioinformatics*. 2025;41(6). doi:10.1093/bioinformatics/btaf283

International Journal of General Medicine

Dovepress
Taylor & Francis Group

Publish your work in this journal

The International Journal of General Medicine is an international, peer-reviewed open-access journal that focuses on general and internal medicine, pathogenesis, epidemiology, diagnosis, monitoring and treatment protocols. The journal is characterized by the rapid reporting of reviews, original research and clinical studies across all disease areas. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/international-journal-of-general-medicine-journal